Gaussian Regularized Sliced Inverse Regression

Caroline Bernard-Michel, Laurent Gardes, Stéphane Girard*

Laboratoire Jean-Kuntzmann & INRIA Rhône-Alpes, team Mistis,
Inovallée, 655, av. de l'Europe, Montbonnot, 38334 Saint-Ismier cedex, France,
(* corresponding author, Stephane.Girard@inrialpes.fr)

Abstract

Sliced Inverse Regression (SIR) is an effective method for dimension reduction in high-dimensional regression problems. The original method, however, requires the inversion of the predictors covariance matrix. In case of collinearity between these predictors or small sample sizes compared to the dimension, the inversion is not possible and a regularization technique has to be used. Our approach is based on a Fisher Lecture given by R.D. Cook where it is shown that SIR axes can be interpreted as solutions of an inverse regression problem. In this paper, a Gaussian prior distribution is introduced on the unknown parameters of the inverse regression problem in order to regularize their estimation. We show that some existing SIR regularizations can enter our framework, which permits a global understanding of these methods. Three new priors are proposed leading to new regularizations of the SIR method. A comparison on simulated data is provided.

Keywords: Inverse regression, regularization, sufficient dimension reduction.

1 Introduction

Many methods have been developed for inferring the conditional distribution of an univariate response Y given a predictor X in $\mathbb{R}^p$, ranging from linear regression [14] to support vector regression [12]. When p is large, sufficient dimension reduction aims at replacing the predictor X by its projection onto a subspace of smaller dimension without loss of information on the conditional distribution of Y given X. In this context, the central subspace, denoted by $S_{Y|X}$ plays an important role. It is defined as the smallest subspace such that, conditionally on the projection of X on $S_{Y|X}$, Y and X are independent. In other words, the projection of X on $S_{Y|X}$ contains all the information on Y that is available in the predictor X.

The estimation of the central subspace has received considerable attention since the past twenty years. Without intending to be exhaustive, we refer to Sliced Inverse Regression (SIR) [21], sliced average variance estimation [7], and graphical regression [8] methods. Among them, SIR seems to be the most popular one. The original method has been adapted to various frameworks and the relative asymptotic properties have been derived, see for instance [20, 32, 24, 18, 25]. We also refer to [3] for a study of the SIR finite sample properties, to [15, 27] for the estimation of $K = \dim(S_{Y|X})$, the dimension of the central subspace, and to [1, 16] for extension to functional covariates.

Assuming that K is known and introducing $\Sigma = \mathrm{Cov}(X)$, in the SIR methodology, a basis of the central subspace is obtained by computing the eigenvectors associated to the largest K eigenvalues of $\Sigma^{-1}\mathrm{Cov}(\mathbb{E}(X|Y))$. Unfortunately, the classical $n-$ sample estimate $\hat{\Sigma}$ of Σ can be singular, or at least ill-conditioned, in several situations. Indeed, since $\mathrm{rank}(\hat{\Sigma}) \leq \min(n-1, p)$, if $n \leq p$ then $\hat{\Sigma}$ is singular. Even when n and p are of the same order, $\hat{\Sigma}$ is ill-conditioned, and its inversion introduces numerical instabilities in the estimation of the central subspace. Similar phenomena occur when the coordinates of X are highly correlated.

Some regularizations of the SIR method have been proposed to overcome this limitation. In [6] and [22], a Principal Component Analysis (PCA) is used as a preprocessing step in order to eliminate the directions in which the random vector X is degenerated. Thus, for a properly chosen dimension d of the projection subspace, the covariance matrix of the projected observations is regular. In the sequel, this technique will be referred to as PCA+SIR. Another method consists in adopting a ridge regression technique (see for instance [14], Chapter 17) *i.e.* replaces the sample estimate $\hat{\Sigma}$ by a perturbed version $\hat{\Sigma} + \tau I_p$ where I_p is the $p \times p$ identity matrix and τ is a positive real number [31]. Here, the idea is that, for τ large enough, $\hat{\Sigma} + \tau I_p$ is regular and its condition number increases with τ. Similarly, in [28, 29], regularized discriminant analysis [17] is adapted to the SIR framework. More recently, it is proposed in [23] to interpret SIR as an optimization problem and to introduce L_1- and L_2- penalty terms in the optimized criterion.

Our approach is based on a Fisher Lecture given by R.D. Cook [10] where it is shown that the axes spanning the central subspace can be interpreted as the solutions of an inverse regression problem. In this paper, a Gaussian prior is introduced on the unknown parameters of the inverse regression problem in order to regularize their estimation. We show that the previously mentioned techniques [6, 22, 31] can enter our framework, which permits a global understanding of these methods. Three new priors are proposed leading to new regularizations of the SIR method. A comparison with the L_2- regularization introduced in [23] is also provided. It is shown that, from the theoretical point of view, the proposed L_2- regularization cannot be justified.

This paper is organized as follows. In Section 2, an adaptation of the inverse regression model to our framework is presented. Section 3 is dedicated to the regularization aspects. Theoretical comparisons with existing approaches as well as new methods are provided. Finite sample properties are illustrated in Section 4. Proofs are postponed to the Appendix.

2 Inverse regression without regularization

Consider X a $\mathbb{R}^p-$ random vector, Y the real response variable and let us denote by $S_{Y|X}$ the central subspace. In the following, for the sake of simplicity, we assume that $K = \dim(S_{Y|X}) = 1$, the case $1 < K < p$ being discussed in Section 3. We thus introduce $\beta \in \mathbb{R}^p$ such that $S_{Y|X} = \text{span}(\beta)$. In Subsection 2.1, the considered inverse regression model is presented. The estimation of the unknown parameters is discussed in Subsection 2.2 and the links with the SIR method are established in Subsection 2.3.

2.1 Single-index inverse regression model

The following inverse single-index regression model is considered (see [10], equation (2) for the multi-index model):

$$X = \mu + c(Y)Vb + \varepsilon, \tag{1}$$

where μ and b are non-random $\mathbb{R}^p-$ vectors, ε is a centered $\mathbb{R}^p-$ Gaussian random vector, independent of Y, with covariance matrix $\text{Cov}(\varepsilon) = V$ and $c : \mathbb{R} \to \mathbb{R}$ is a nonrandom function. Under this model,

$$\mathbb{E}(X|Y = y) = \mu + c(y)Vb,$$

and thus, after translation by μ, the conditional expectation of X given Y is a degenerated random vector located in the direction Vb. From [10], Proposition 1, b corresponds to the direction β of the central subspace. In the sequel, it will appear that, under appropriate conditions, the maximum likelihood estimator of b is (up to a scale parameter) the SIR estimator of β. Moreover, note that, under (1), one has

$$c(y) = \frac{\mathbb{E}(b^t(X - \mu)|Y = y)}{b^tVb}. \tag{2}$$

Now, restricting ourselves to the single-index case, the forward model of SIR asserts that there exists a univariate link function g such that $\mathbb{E}(Y|X) = g(b^tX)$ or equivalently, $b^tX = g^{-1}(\mathbb{E}(Y|X))$. Thus, replacing in (2) yields

$$c(y) = \frac{g^{-1}(y) - b^t\mu}{b^tVb},$$

i.e. the coordinate function is, up to an affine function, the inverse of the link function in the single-index forward model of SIR.

2.2 Maximum likelihood estimation

We now address the estimation of the coordinate function $c(.)$, the direction b, the covariance matrix V and the location parameter μ in model (1). To this end, we focus on a projection estimator of the unknown function $c(.)$. More precisely, it is expanded as a linear combination of h basis functions $s_j(.)$, $j = 1, \dots, h$:

$$c(.) = \sum_{j=1}^{h} c_j s_j(.), \tag{3}$$

where the coefficients c_j, $j = 1, \dots, h$ are unknown whereas h is supposed to be known. Introducing $c = (c_1, \dots, c_h)^t$ and $s(.) = (s_1(.), \dots, s_h(.))^t$, model (1) can be rewritten as

$$X = \mu + s^t(Y)cVb + \varepsilon, \;\; \varepsilon \sim \mathcal{N}(0, V), \tag{4}$$

where $\mathcal{N}(0, V)$ is the multivariate centered Gaussian distribution with covariance matrix V. In the sequel we denote by

$$\rho = \frac{b^t \Sigma b - b^t V b}{b^t V b},$$

the signal to noise ratio in the direction b. Let (X_i, Y_i), $i = 1, \dots, n$ be a sample of independent random variables distributed as (X, Y). Clearly, estimating (μ, V, b, c) by maximization of the likelihood in model (4) consists in minimizing

$$G(\mu, V, b, c) = \log\det V + \frac{1}{n}\sum_{i=1}^{n}(\mu + s^t(Y_i)cVb - X_i)^t V^{-1}(\mu + s^t(Y_i)cVb - X_i), \tag{5}$$

with respect to (μ, V, b, c). Note that $G(\mu, V, b, c)$ can also be interpreted as a discrepancy functional, see equation (5) in [11]. Up to our knowledge, the introduction of such functional in the inverse regression framework is due to [9]. Let us introduce the $h \times h$ empirical covariance matrix W of $s(Y)$ defined by

$$W = \frac{1}{n}\sum_{i=1}^{n}(s(Y_i) - \bar{s})(s(Y_i) - \bar{s})^t,$$

the $h \times p$ matrix M defined by

$$M = \frac{1}{n}\sum_{i=1}^{n}(s(Y_i) - \bar{s})(X_i - \bar{X})^t,$$

and $\hat{\Sigma}$ the empirical $p \times p$ covariance matrix of X

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^t,$$

where

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \;\text{ and }\; \bar{s} = \frac{1}{n}\sum_{i=1}^{n} s(Y_i).$$

Lemma 1 *Using the above notations, $G(\mu, V, b, c)$ can be rewritten as*

$$\begin{aligned} G(\mu, V, b, c) &= \log\det V + tr(\hat{\Sigma}V^{-1}) + (\mu - \bar{X} + \bar{s}^t cVb)^t V^{-1}(\mu - \bar{X} + \bar{s}^t cVb) \\ &+ (c^t Wc)(b^t Vb) - 2c^t Mb. \end{aligned}$$

The maximum likelihood estimators of μ, V, b and c are closed-form.

Proposition 1 *Under (4), if W and $\hat{\Sigma}$ are regular, then the maximum likelihood estimator of (μ, V, b, c) is defined by:*

- $\hat{b}$ *is the eigenvector associated to the largest eigenvalue* $\hat{\lambda}$ *of* $\hat{\Sigma}^{-1}M^tW^{-1}M$,
- $\hat{c} = \dfrac{1}{\hat{b}^t\hat{V}\hat{b}}W^{-1}M\hat{b}$,
- $\hat{\mu} = \bar{X} - \bar{s}^t\hat{c}\hat{V}\hat{b}$,
- $\hat{V} = \hat{\Sigma} - \dfrac{\hat{\lambda}}{\hat{b}^t\hat{\Sigma}\hat{b}}\hat{\Sigma}\hat{b}\hat{b}^t\hat{\Sigma}$.

As a consequence of the above equation, one also has $\hat{\lambda} = 1 - \hat{b}^t\hat{V}\hat{b}/\hat{b}^t\hat{\Sigma}\hat{b}$, which provides an estimation of the signal to noise ratio in the direction b through

$$\hat{\rho} = \hat{\lambda}/(1 - \hat{\lambda}). \tag{6}$$

Let us now show that the SIR method corresponds to the particular case of piecewise constant basis functions $s_j(.)$, $j = 1, \ldots, h$.

2.3 Sliced Inverse Regression (SIR)

Suppose the range of Y is partitioned into $h+1$ non-overlapping slices S_j, $j = 1, \ldots, h+1$ and consider the h basis functions defined by

$$s_j(.) = \mathbb{I}\{. \in S_j\}, \;\; j = 1, \ldots, h \tag{7}$$

where $\mathbb{I}$ is the indicator function. Let us denote by n_j the number of Y_i in slice $j = 1, \ldots, h+1$, define the corresponding proportion by $f_j = n_j/n$, the empirical mean of X given $Y \in S_j$ by

$$\bar{X}_j = \frac{1}{n_j}\sum_{Y_i \in S_j} X_i$$

and let $\hat{\Gamma}$ be the $p \times p$ empirical "between slices" covariance matrix defined by

$$\hat{\Gamma} = \sum_{j=1}^{h+1} f_j(\bar{X}_j - \bar{X})(\bar{X}_j - \bar{X})^t.$$

In this context, the following consequence of Proposition 1 can be established.

Corollary 1 *Under (4) and (7), if $\hat{\Sigma}$ is regular, then the maximum likelihood estimator $\hat{b}$ of b is the eigenvector associated to the largest eigenvalue of $\hat{\Sigma}^{-1}\hat{\Gamma}$.*

To summarize, the maximum likelihood estimator of b is (up to a scale factor) the classical SIR estimator of the direction β spanning the central subspace. The next section is dedicated to the introduction of a regularization in the inverse regression problem in order to avoid the inversion of $\hat{\Sigma}$.

3 Regularized inverse regression

First, we present in Subsection 3.1 how the introduction of a prior information on the unknown direction b can overcome the SIR limitations due to the ill-conditioning or singularity of $\hat{\Sigma}$. Second, some links with the existing SIR regularizations are highlighted in Subsection 3.2. Finally, basing on our framework, three new regularizations of the SIR method are introduced in Subsection 3.3.

3.1 Introducing a Gaussian prior

We propose to introduce a prior information on the projection of X on b appearing in the inverse regression model. More precisely, we focus on

$$\frac{\mathbb{E}(b^t(X-\mu)|Y)b}{b^tVb} = c(Y)b = s^t(Y)cb,$$

from (2) and (3), assuming that

$$(1+\rho)^{-1/2}\,(s(Y)-\bar{s})^t cb \sim \mathcal{N}(0,\Omega). \tag{8}$$

The role of the matrix Ω is to describe which directions in $\mathbb{R}^p$ are the most likely to contain b. Some examples are provided in the next two paragraphs. The scalar $(1+\rho)^{-1/2}$ is introduced for normalization purposes, permitting to preserve the interpretation of the eigenvalue in terms of signal to noise ratio. As a comparison, in [2], a Bayesian estimation method is proposed basing on B-splines approximation of the link function g in the forward model and a Fisher-von Mises prior on the direction b. In our approach, working on the inverse regression model allows to obtain explicit solutions, see Lemma 2 and Proposition 2 below.

Lemma 2 *Maximum likelihood estimators are obtained by minimizing*

$$G_\Omega(\mu,V,b,c) = G(\mu,V,b,c) + \frac{(b^t\Omega^{-1}b)(b^tVb)(c^tWc)}{b^t\Sigma b} \tag{9}$$

with respect to (μ,V,b,c).

Comparing to (5), the additional term due to the prior information can be read as a regularization term in Tikhonov theory, see for instance [30], Chapter 1, penalizing large projections. The following result can be stated:

Proposition 2 *Under (4) and (8), if W and $\Omega\hat{\Sigma} + I_p$ are regular, then the maximum likelihood estimator of (μ, V, b, c) is defined by:*

- *$\hat{b}$ is the eigenvector associated to the largest eigenvalue $\hat{\lambda}$ of $(\Omega\hat{\Sigma}+I_p)^{-1}\Omega M^t W^{-1} M$,*
- $\hat{c} = \dfrac{1}{(1+\eta(\hat{b}))\hat{b}^t\hat{V}\hat{b}} W^{-1}M\hat{b}$, *where* $\eta(\hat{b}) = \dfrac{\hat{b}^t\Omega^{-1}\hat{b}}{\hat{b}^t\hat{\Sigma}\hat{b}}$,
- $\hat{\mu} = \bar{X} - \bar{s}^t\hat{c}\hat{V}\hat{b}$,
- $\hat{V} = \hat{\Sigma} - \dfrac{\hat{\lambda}}{\hat{b}^t\hat{\Sigma}\hat{b}}\hat{\Sigma}\hat{b}\hat{b}^t\hat{\Sigma}$.

Compared to Proposition 1, the inversion of $\hat{\Sigma}$ is replaced by the inversion of $\Omega\hat{\Sigma} + I_p$. Thus, for a properly chosen prior matrix Ω, the numerical instabilities in the estimation of b disappear. Note that, since the estimation of V is formally the same as in Proposition 1, the interpretation (6) of $\hat{\lambda}$ still holds. As previously, this result can be applied to the particular case of the SIR method.

Corollary 2 *Under (4), (7) and (8), if $\Omega\hat{\Sigma} + I_p$ is regular, then the maximum likelihood estimator of b is the eigenvector associated to the largest eigenvalue of $(\Omega\hat{\Sigma} + I_p)^{-1}\Omega\hat{\Gamma}$.*

In the following, the above estimator of the direction b will be referred to as the Gaussian Regularized Sliced Inverse Regression (GRSIR) estimator. Let us emphasize that the GRSIR estimator can be extended to the multi-index situation by considering the eigenvectors $b_1, \dots, b_K$ associated to the K largest eigenvalues $\lambda_1 > \cdots > \lambda_K$ of $(\Omega\hat{\Sigma} + I_p)^{-1}\Omega\hat{\Gamma}$. For instance, one can show that b_2 maximizes G_Ω under the orthogonality constraint $b_1^t(\Sigma + \Omega^{-1})b_2 = 0$.

Some examples of possible prior covariance matrices Ω are now presented.

3.2 Links with existing methods

In all the next examples, a non-negative parameter τ is introduced in the prior covariance matrix in order to tune the importance of the penalty term in (9). Consequently, in the sequel, τ is called a regularization parameter.

Classical SIR approach. It is easily seen from Corollary 2 that choosing the prior covariance matrix $\Omega_0 = (\tau\hat{\Sigma})^{-1}$ in GRSIR gives back SIR, and this for all $\tau > 0$. This prior matrix indicates that directions corresponding to small variances are most likely, *i.e* the SIR method favors directions in which $\hat{\Sigma}$ is close to singularity. In practice, this choice yields instabilities in the estimation.

Ridge approach [28, 29, 31]. The simplest choice for the prior covariance matrix is $\Omega_1 = \tau^{-1} I_p$. In this case, the identity matrix indicates that no privileged direction for b is available. Following Corollary 2, the corresponding GRSIR estimator of b is the eigenvector of $(\hat{\Sigma}+\tau I_p)^{-1}\hat{\Gamma}$ associated to its largest eigenvalue, which is the ridge estimator introduced independently in [31] and [28, 29].

PCA+SIR approach [6, 22]. As already seen, a popular technique to overcome the singularity problems of $\hat{\Sigma}$ is to use PCA as a preprocessing step [6, 22]. The principle is the following. Let $d \in \{1, \dots, p\}$ be fixed and denote by $\lambda_1 \geq \dots \geq \lambda_d$ the d largest eigenvalues of $\hat{\Sigma}$ (supposed to be positive), $q_1, \dots, q_d$ the associated eigenvectors and $S_d = \text{span}(q_1, \dots, q_d)$ the linear subspace spanned by $q_1, \dots, q_d$. The first step consists in projecting the predictors on S_d. The second step is to perform SIR in this subspace. The next result shows that this method corresponds to a particular prior covariance matrix.

Proposition 3 *PCA+SIR corresponds to GRSIR with prior covariance matrix*

$$\Omega_2 = \frac{1}{\tau} \sum_{j=1}^{d} \frac{1}{\lambda_j} q_j q_j^t,$$

where $\tau > 0$ is arbitrary.

Let us note that although Ω_2 depends on τ, the GRSIR estimator does not, since there is no regularization parameter in the PCA+SIR methodology.

Li and Yin's approach [23]. Their proposed L_2- regularization consists in estimating (b, c) by minimization of

$$H_\tau(b, c) = \sum_{j=1}^{h} f_j (\mu + \hat{\Sigma} c_j b - \bar{X}_j)^t N (\mu + \hat{\Sigma} c_j b - \bar{X}_j) + \tau b^t b,$$

the matrix N being either $N = I_p$ or $N = \hat{\Sigma}^{-1}$. In our opinion, this approach suffers from a lack of invariance since the functional H_τ does not penalize the same way two different axes (b and $2b$ for instance) defining the same direction. As a consequence, we have shown ([5], Proposition 1) that the only possible solution $\hat{b}$ of the minimization problem is $\hat{b} = 0$: In view of this result, the proposed alternating least squares algorithm ([23], Section 2) cannot be justified theoretically. As a comparison, our method does not yield this kind of problem thanks to the invariance property: $G_\Omega(\mu, V, tb, c/t) = G_\Omega(\mu, V, b, c)$ for all real numbers $t \neq 0$. We now propose some alternative choices of the covariance matrix Ω yielding new regularizations of the SIR method.

3.3 Three new SIR regularizations

Tikhonov regularization. An alternative choice of the prior covariance matrix is $\Omega_3 = \tau^{-1}\hat{\Sigma}$. Comparing Ω_3 to the matrix Ω_0 associated to the SIR method, it appears that the underlying ideas are opposite. Here, directions corresponding to large variances are most likely. The associated GRSIR estimator of the direction b is the eigenvector of $(\hat{\Sigma}^2+\tau I_p)^{-1}\hat{\Sigma}\hat{\Gamma}$ associated to its largest eigenvalue. In the following, this estimator will be referred to as the Tikhonov estimator. Indeed, let us recall that the classical SIR estimator is obtained by a spectral decomposition of $\hat{\Sigma}^{-1}\hat{\Gamma}$. For all $k=1,\dots,p$, denote by x_k the $k-$th column of this matrix. Computing x_k is equivalent to solving with respect to x the linear system $\hat{\Sigma}x=\hat{\Gamma}_k$ where $\hat{\Gamma}_k$ is the $k-$th column of $\hat{\Gamma}$. The associated Tikhonov minimization problem (see (1.34) in [30]) can be written as

$$x_k = \arg\min_x \|\hat{\Sigma}x-\hat{\Gamma}_k\|^2+\tau\|x\|^2=(\hat{\Sigma}^2+\tau I_p)^{-1}\hat{\Sigma}\hat{\Gamma}_k.$$

Thus, in this framework, $(\hat{\Sigma}^{-1}\hat{\Gamma})_k$ is estimated by $(\hat{\Sigma}^2+\tau I_p)^{-1}\hat{\Sigma}\hat{\Gamma}_k$ and consequently $\hat{\Sigma}^{-1}\hat{\Gamma}$ is estimated by $(\hat{\Sigma}^2+\tau I_p)^{-1}\hat{\Sigma}\hat{\Gamma}$.

Dimension reduction approaches. It has been seen in Proposition 3, that the PCA+SIR approach is equivalent to using the prior covariance matrix Ω_2 in GRSIR. The following result is an extension to more general covariance matrices.

Proposition 4 *For all real function φ let*

$$\Omega(\varphi)=\sum_{j=1}^{d}\varphi(\lambda_j)q_jq_j^t.$$

Then, the associated GRSIR estimator can be obtained by first projecting the predictors on $S_d=span(q_1,\dots,q_d)$ and second performing GRSIR on the projected predictors with prior covariance matrix

$$\tilde{\Omega}(\varphi)=\sum_{j=1}^{p}\varphi(\lambda_j)q_jq_j^t.$$

The dimension d plays the role of a "cut-off" parameter, since when computing $\hat{b}$, all directions $q_{d+1},\dots,q_p$ are discarded. Three illustrations of this result can be given:

- Choosing $\varphi(t)=1/(\tau t)$, we obtain $\Omega(1/(\tau\text{Id}))=\Omega_2$ and $\tilde{\Omega}(1/(\tau\text{Id}))=(\tau\hat{\Sigma})^{-1}=\Omega_0$, where Id is the identity function. It appears that Proposition 3 is a particular case of Proposition 4. As already discussed, since the choice of Ω_0 as a prior covariance matrix seems not very natural, we thus propose two new choices.

- First, $\varphi(t)=1/\tau$ yields

$$\Omega_4\overset{def}{=}\Omega(1/\tau)=\frac{1}{\tau}\sum_{j=1}^{d}q_jq_j^t,$$

  and $\tilde{\Omega}(1/\tau)=I_p/\tau=\Omega_1$. Consequently, this new method consists in applying the ridge approach [31] on the projected predictors, the interpretation being that, in the subspace S_d, all directions share the same prior probability. This method will be referred to as PCA+ridge.

- Second, $\varphi(t) = t/\tau$ yields

  $$\Omega_5 \overset{def}{=} \Omega(\mathrm{Id}/\tau) = \frac{1}{\tau}\sum_{j=1}^{d} \lambda_j q_j q_j^t,$$

  and $\tilde{\Omega}(\mathrm{Id}/\tau) = \hat{\Sigma}/\tau = \Omega_3$. This new method consists in applying Tikhonov approach on the projected predictors. In this context, directions of S_d carrying a large fraction of the total variance of X are more likely. This method will be referred to as PCA+Tikhonov.

4 Numerical experiments

GRSIR methods associated to the prior covariance matrices Ω_0 (SIR), Ω_1 (ridge), Ω_2 (PCA+SIR), Ω_3 (Tikhonov), Ω_4 (PCA+ridge) and Ω_5 (PCA+Tikhonov) are compared on simulated data. A sample of size $n = 100$ of the random pair (X, Y) is considered, where $X \in \mathbb{R}^p$ with $p = 50$ and $Y \in \mathbb{R}$. The random vector X is Gaussian, centered, with covariance matrix $\Sigma = Q\Delta Q^t$ where Δ is the diagonal matrix containing the eigenvalues of Σ defined by Δ =diag$\{p^\theta, (p-1)^\theta, \dots, 1^\theta\}$ and Q is a randomly chosen orthogonal matrix. Several values of θ will be considered. Note that the condition number of Σ is given by p^θ and is thus an increasing function of θ. As suggested by [19], the orthogonal matrix Q is obtained as follow: First, we construct a $p \times p$ matrix where each coefficients are randomly chosen from a Gaussian distribution. Next, the matrix Q is obtained by applying the QR-decomposition on this matrix. We consider two SIR models:

$$\textbf{Model 1} \quad Y = \sin\left(\frac{\pi}{2\sigma}\beta^t X\right) + \varepsilon,$$

$$\textbf{Model 2} \quad Y = \left|\frac{\beta^t X}{\sigma} - 1/2\right| + \varepsilon,$$

where σ is the standard deviation of the projection of X on β *i.e.* $\sigma = (\beta^t \Sigma \beta)^{1/2}$, ε is a centered Gaussian random value with standard deviation 0.03 independent of X. The true index is $\beta = 5^{-1/2} Q(1,1,1,1,1,0,\dots,0)^t$. Hence in the direction of β, the variance of X is large.

In all the experiments described below, we replicate $N = 100$ times Models 1 and 2. More precisely, for Model 1, we compute the pairs $(X, Y^{(r)})$, $r = 1, \dots, N$ where

$$Y^{(r)} = \sin\left(\frac{\pi}{2\sigma}\beta^t X\right) + \varepsilon_r,\ \varepsilon_r \sim \mathcal{N}(0, (0.03)^2),$$

with ε_r independent of X. The same is done for Model 2. We then obtained N estimators $\hat{b}^{(r)}$, $r = 1, \dots, N$ of β. In order to evaluate the quality of the estimate $\hat{b}$, we compute two criteria. The first one is the mean of the squared cosine (MSC):

$$\mathrm{MSC} = \frac{1}{N}\sum_{r=1}^{N} (\beta^t \hat{b}^{(r)})^2.$$

The second one is a kind of variance of the squared cosine (VSC):

$$\text{VSC} = \frac{1}{N(N-1)} \sum_{r=1}^{N} \sum_{r \neq s} ((\hat{b}^{(s)})^t \hat{b}^{(r)})^2.$$

These criteria can be adapted to multi-index models [15]. The closer these quantities are to 1, the better the estimation is.

First experiment: We illustrate the dependence of the above methods with respect to the regularization parameter τ. In this aim, we compute the two previous defined criteria as function of the regularization parameter. A logarithmic scale was adopted, 150 values of $\log(\tau)$ regularly distributed in $[-5, 25]$ were considered. Here, we limit ourselves to $\theta = 2$. Moreover, we choose $d = 20$ in the PCA+SIR, PCA+ridge and PCA+Tikhonov methods. Results are displayed on Figure 1 and Figure 2 for Model 1 and on Figure 3 and Figure 4 for Model 2. It appears that the classical SIR approach gives very poor results in such a situation where n and p are of the same order ($n/p = 2$). Ridge and Tikhonov regularizations can bring a significant improvement provided τ is large enough. The VSC criterion of the ridge and the Tikhonov methods is better than the one of the classical SIR method if the regularization parameter is taken sufficiently large. PCA+SIR obtains reasonable results compared to SIR, with the advantage of do not requiring the selection of τ. The selection of d is addressed in our third experiment. Note that PCA+ridge and PCA+Tikhonov methods are less sensitive to the choice of τ than ridge and Tikhonov methods. Concerning the criterion VSC, PCA+ridge and PCA+Tikhonov methods both outperform the PCA+SIR for sufficiently large values of the regularization parameter. For one of the $N = 100$ simulated datasets, the pairs $(\beta^t X_i, \hat{b} X_i)$, $i = 1, \ldots, n$ are represented on Figure 5 for Model 1 and on Figure 6 for Model 2. In order to make a comparison, the estimated axis $\hat{b}$ is computed with the classical SIR method and the PCA+Tikhonov method. It appears clearly that the SIR method leads to very bad oriented estimator of β which is not the case with the PCA+Tikhonov method.

Second experiment: The robustness with respect to the condition number is investigated by varying θ in $\{0, 0.1, 0.2, \ldots, 3\}$. For each value within this set, the optimal regularization parameter has been selected for each method and the corresponding MSC criterion is displayed on Figure 7 and Figure 8. This is done only for Model 1. Clearly, the classical SIR method is very sensitive to the ill-conditioning of the covariance matrix. For all the other considered methods, results are getting better while the condition number increases. Note that ridge and Tikhonov methods as well as PCA+ridge and PCA+Tikhonov yield very similar results.

Third experiment: Illustration of the role of d in PCA+SIR, PCA+ridge and PCA+Tikhonov methods. Here, the condition number is fixed by choosing $\theta = 2$. For each value of d in $\{0, 1, \ldots, p\}$, the optimal regularization parameter has been selected for each method and the corresponding MSC criterion is displayed on Figure 9. Only the Model 1 is considered here. One can see that the PCA+SIR method is very sensitive to d. Indeed, if d is large, then this approach reduces to SIR, whose accuracy is low for large dimensions. At the opposite, PCA+ridge and PCA+Tikhonov results remain stable as d increases, since these methods get close to ridge and Tikhonov methods respectively.

5 Retrieval of Mars surface physical properties from hyperspectral images

We propose here to apply GRSIR in the context of a nonlinear inverse problem in remote sensing. Hyperspectral remote sensing is a promising space technology regularly selected by agencies with regard to the exploration and observation of planets. It allows to collect for each pixel of a scene, the intensity of light energy reflected from materials as it varies across different wavelengths. Hundreds spectels in the visible and near infra-red are recorded, making it possible to observe a continuous spectrum for each image cell. The analyze of these spectral signatures allows to identify the physical, chemical or mineralogical properties of the surface and of the atmosphere that may help to understand the geological and climatological history of planets.

Our goal is to evaluate the physical properties of surface materials on Mars planet from hyperspectral images collected by the OMEGA instrument aboard Mars express spacecraft. The used approach is based on the estimation of the functional relationship G between some physical parameters Y and observed spectra X. For this purpose, a database of synthetic spectra is generated by a physical radiative transfer model and used to estimate G. The high dimension of spectra ($p = 184$ wavelengths) is reduced by using GRSIR. Results are compared with the traditional SIR approach.

5.1 Data

In this application, we focus on an observation of the south pole of Mars at the end of summer. It has been collected during orbit 61 by the French imaging spectrometer OMEGA on board Mars Express Mission. A detailed analysis of this image [13] revealed that this portion of Mars mainly contains water ice, carbon dioxide and dust. This has led to the physical modeling of individual spectra with a surface reflectance model. This model allows the generation of 12.000 synthetic spectra with the corresponding parameters that constitute a learning database. Here, we focus on the terrain unit of strong CO_2 concentration determined by a classification method based on wavelets [26]. It contains approximately 9,000 spectra to reverse. The 5 most important parameters characterizing the morphology of these spectra are the proportions of CO_2 ice, water ice and dust; and the grain sizes of water ice and CO_2 ice. In the sequel only one parameter, the grain size of CO_2 ice is presented. A detailed analysis for all other parameters can be found in [4].

5.2 Methodology

Two methods are compared in order to reverse the hyperspectral image: SIR and GRSIR with the PCA + ridge approach. Both methods allow to choose a lower dimensional regressor space, sufficient enough to describe the relationship of interest. In practice, it appears that a unidimensional regressor space gives satisfactory results [4]. Thus, we assume that there exists a function g from $\mathbb{R}$ to $\mathbb{R}$ such that: $G(X) = g(< \beta, X >)$. As a consequence, the estimation of the relationship G reduces to computing one direction β, the univariate function g being estimated by piecewise linear regression. In the PCA+ridge approach, the cut-off dimension is fixed to $d = 20$. The regularization parameter, fixed to $\tau = 10^{-4.2}$, is chosen to minimize the mean squared error when estimating the grain size of CO_2 ice of the learning database itself. An example of the relationship between reduced spectra and the grain size of CO_2 estimated using GRSIR is presented on Figure 10. It shows that the relationship is nonlinear and that one direction seems to be sufficient to estimate the grain size of CO_2.

5.3 Results

With GRSIR methodology, the inversion of the image from orbit 61 shows a smooth mapping of the grain size of CO_2 (see Figure 12) making it possible to distinguish some areas with great sizes of CO_2 ice on the boundaries and some areas with small values inside the cap. On the opposite, with the traditional SIR approach, estimates assume a small number of values that seem to be distributed randomly and correspond to the minimum and maximum values of the parameter in the learning database (see Figure 11). These poor results can be explained by the very high condition number, about 10^{14}, of the empirical covariance matrix. Since no ground data is available, it is quite difficult to quantify the accuracy of GRSIR estimates. However, comparisons with other approaches or with estimates from other hyperspectral images from the same portion of Mars [4] give consistent results. GRSIR approach then appears promising for model inversion in remote sensing.

6 Concluding remarks

A new framework has been presented to regularize the SIR method. It provides new interpretations of the existing methods as well as the construction of new regularization techniques. Among them, it appears that the PCA+ridge and PCA+Tikhonov methods are interesting alternatives to the existing PCA+SIR [6, 22] and ridge [31] methods. The use of a "cut-off" dimension d permits to limit the sensitivity to the choice of the regularization parameter τ. The choice of this dimension itself seems not crucial since for large value the above methods are close to the ridge and Tikhonov techniques. In our experiments, the choice $d = p/2$ appears as a good heuristics in most situations. Of course, an automatic selection of (τ, d) would be of interest. To this end, the construction of a generalized cross-validation criterion is under investigation. We also plan to study the introduction on non-Gaussian priors in order to obtain L_1- penalizations and thus sparse estimates of β.

References

[1] Amato, U., Antoniadis, A. and De Feiss, I. (2006). Dimension reduction in functional regression with applications. *Computational Statistics and Data Analysis*, **50**, 2422–2446.

[2] Antoniadis, A., Grégoire, G. and McKeague, I.W. (2004). Bayesian estimation in single-index models. *Statistica Sinica*, **14**, 1147–1164.

[3] Aragon, Y. and Saracco, J. (1997). Sliced Inverse Regression: An appraisal of small sample alternatives to slicing. *Computational Statistics*, **12**, 109–130.

[4] Bernard-Michel, C., Gardes, L. and Girard, S. (2007). Estimation of Mars surface physical properties from hyperspectral images using Sliced Inverse Regression. *Technical report*, INRIA, `http://hal.inria.fr/inria-00187444`.

[5] Bernard-Michel, C., Gardes, L. and Girard, S. (2008). A note on Sliced Inverse Regression with regularizations. *Biometrics*, to appear.

[6] Chiaromonte, F. and Martinelli, J. (2002). Dimension reduction strategies for analyzing global gene expression data with a response. *Mathematical Biosciences*, **176**, 123–144.

[7] Cook, R.D. and Weisberg, S. (1991). Discussion of "Sliced Inverse Regression for dimension reduction" by K.C. Li, *Journal of the American Statistical Association*, **86**, 328–332.

[8] Cook, R.D. (1998). *Regression graphics. Ideas for studying regressions through graphics.* Wiley Series in Probability and Statistics, New York.

[9] Cook, R.D. (2004). Testing predictor contributions in sufficient dimension reduction. *The Annals of Statistics*, **32**, 1062–1092.

[10] Cook, R.D. (2007). Fisher Lecture: Dimension Reduction in Regression. *Statistical Science*, **22**(1), 1–26.

[11] Cook, R.D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of the American Statistical Association*, **100**, 410–428.

[12] Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.* Cambridge University Press.

[13] Douté, S., Schmitt, B., Langevin, Y., Bibring, J-P., Altieri, F., Bellucci, G., Gondet, B. and Poulet, F. (2007). South Pole of Mars: Nature and composition of the icy terrains from Mars Express OMEGA observations. *Planetary and Space Science*, **55**(1-2), 113–133.

[14] Draper, N.R. and Smith, R. (1998). *Applied regression analysis (3rd edition)*, Wiley.

[15] Ferré, L. (1998). Determining the dimension in Sliced Inverse Regression and related methods. *Journal of the American Statistical Association*, **93**, 132–140.

[16] Ferré, L. and Yao, A.F. (2005). Smoothed functional inverse regression. *Statistica Sinica*, **15**, 665–683.

[17] Friedman, J.H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, **84**, 165–175.

[18] Gannoun, A. and Saracco, J. (2003). An asymptotic theory for SIR$_\alpha$ method. *Statistica Sinica*, **13**, 297–310.

[19] Heiberger, R. (1978). Generation of random orthogonal matrices. *Journal of the Royal Statistical Society, Series C*, **27**, 199–206.

[20] Hsing, T. and Carroll, R.J. (1992). An asymptotic theory for Sliced Inverse Regression. *The Annals of Statistics*, **20**(2), 1040–1061.

[21] Li, K.C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, **86**, 316–327.

[22] Li, L. and Li, H. (2004). Dimension reduction methods for micro-arrays with application to censored survival data. *Bioinformatics*, **20**(18), 3406–3412.

[23] Li, L. and Yin, X. (2008). Sliced Inverse Regression with regularizations. *Biometrics*, to appear, doi: 10.1111/j.1541-0420.2007.00836.x.

[24] Saracco, J. (1997). An asymptotic theory for Sliced Inverse Regression. *Communications in Statistics, Theory and Methods*, **26**, 2141–2171.

[25] Saracco, J. (2005). Asymptotics for pooled marginal slicing estimator based on SIR$_\alpha$ approach. *Journal of Multivariate Analysis*, **96**(1), 117–135.

[26] Schmidt, F., Douté, S. and Schmitt B. (2007). WAVANGLET: An efficient supervised classifier for hyperspectral images. *Geoscience and Remote Sensing, IEEE Transactions*, **45**(5), 1374-1385.

[27] Schott, J.R. (1994). Determining the dimensionality in Sliced Inverse Regression. *Journal of the American Statistical Association*, **89**, 141–148.

[28] Scrucca, L. (2006). Regularized Sliced Inverse Regression with applications in classification. In Zani S., Cerioli A., Riani M. and Vichi M., editors, *Data Analysis, Classification and the Forward Search*, pp. 59-66, Berlin, Springer-Verlag,

[29] Scrucca, L. (2007). Class prediction and gene selection for DNA microarrays using regularized Sliced Inverse Regression. *Computational Statistics and Data Analysis*, **52**, 438–451.

[30] Vogel, C.R. (2002). *Computational methods for inverse problems*, Society for Industrial and Applied Mathematics, Philadelphia.

[31] Zhong, W., Zeng, P., Ma, P., Liu, J.S. and Zhu, Y. (2005). RSIR: Regularized Sliced Inverse Regression for motif discovery. *Bioinformatics*, **21**(22), 4169–4175.

[32] Zhu, L.X. and Ng, K.W. (1995). Asymptotics of Sliced Inverse Regression. *Statistica Sinica*, **5**, 727–736.

Acknowledgments

This research is partially supported by IAP research network grant nr. P6/03 of the Belgian government (Belgian Science Policy). We are grateful to Sylvain Douté for his important contribution to the data and to the referees whose remarks led to an important improvement of the presentation.

Appendix: Proofs

Proof of Lemma 1. Let us remark that

$$\Delta := G(\mu, V, b, c) - \log\det V = \frac{1}{n}\sum_{i=1}^{n} Z_i^t V^{-1} Z_i \tag{10}$$

where we have defined for $i = 1, \dots, n$,

$$\begin{aligned} Z_i &= \mu + s^t(Y_i)cVb - X_i \\ &= (\mu - \bar{X} + \bar{s}^t cVb) + (s(Y_i) - \bar{s})^t cVb - (X_i - \bar{X}) \\ &:= Z_{1,i} + Z_{2,i} - Z_{3,i}. \end{aligned}$$

Since Z_2 and Z_3 are centered, replacing the previous expansion in (10) yields

$$\Delta = \frac{1}{n}\sum_{i=1}^{n} Z_{1,i}^t V^{-1} Z_{1,i} + \frac{1}{n}\sum_{i=1}^{n} Z_{2,i}^t V^{-1} Z_{2,i} + \frac{1}{n}\sum_{i=1}^{n} Z_{3,i}^t V^{-1} Z_{3,i} - \frac{2}{n}\sum_{i=1}^{n} Z_{2,i}^t V^{-1} Z_{3,i},$$

where

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} Z_{1,i}^t V^{-1} Z_{1,i} &= (\mu - \bar{X} + \bar{s}^t cVb)^t V^{-1}(\mu - \bar{X} + \bar{s}^t cVb) \\
\frac{1}{n}\sum_{i=1}^{n} Z_{2,i}^t V^{-1} Z_{2,i} &= (c^t W c)(b^t V b) \\
\frac{1}{n}\sum_{i=1}^{n} Z_{3,i}^t V^{-1} Z_{3,i} &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{tr}((X_i - \bar{X})^t V^{-1}(X_i - \bar{X})) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{tr}(V^{-1}(X_i - \bar{X})(X_i - \bar{X})^t) \\
&= \mathrm{tr}(V^{-1}\hat{\Sigma}) \\
\frac{2}{n}\sum_{i=1}^{n} Z_{2,i}^t V^{-1} Z_{3,i} &= 2c^t M b,
\end{aligned}$$

and the conclusion follows. ■

Proof of Proposition 1. Annulling the gradients of $G(\mu, V, b, c)$ yields the system of equations

$$\begin{aligned}
\frac{1}{2}\nabla_\mu G &= \hat{V}^{-1}(\hat{\mu} - \bar{X} + \bar{s}^t \hat{c}\hat{V}\hat{b}) = 0, && (11) \\
\frac{1}{2}\nabla_b G &= \left((\hat{c}^t\bar{s})^2 + \hat{c}^t W \hat{c}\right)\hat{V}\hat{b} - M^t\hat{c} + (\hat{c}^t\bar{s})(\hat{\mu} - \bar{X}) = 0, && (12) \\
\frac{1}{2}\nabla_c G &= (\hat{b}^t\hat{V}\hat{b})(\bar{s}\bar{s}^t + W)\hat{c} - M\hat{b} + (\hat{\mu} - \bar{X})^t\hat{b}\bar{s} = 0, && (13) \\
\nabla_V G &= \hat{V}^{-1} + \hat{b}\hat{b}^t\left((\hat{c}^t\bar{s})^2 + \hat{c}^t W\hat{c}\right) - \hat{V}^{-1}\left((\hat{\mu} - \bar{X})(\hat{\mu} - \bar{X})^t + \hat{\Sigma}\right)\hat{V}^{-1} = 0. && (14)
\end{aligned}$$

From (11), we have

$$\hat{\mu} = \bar{X} - \bar{s}^t\hat{c}\hat{V}\hat{b}. \quad (15)$$

Replacing in (12) and (13) yields the simplified system of equations

$$\begin{aligned}
(\hat{c}^t W\hat{c})\hat{V}\hat{b} &= M^t\hat{c}, && (16) \\
(\hat{b}^t\hat{V}\hat{b})W\hat{c} &= M\hat{b}. && (17)
\end{aligned}$$

Assuming W is regular, equation (17) entails

$$\hat{c} = W^{-1}M\hat{b}/\hat{b}^t\hat{V}\hat{b}$$

and replacing in (16) yields

$$(\hat{c}^t W\hat{c})(\hat{b}^t\hat{V}\hat{b})\hat{V}\hat{b} = M^t W^{-1} M\hat{b}. \quad (18)$$

Now, multiplying (14) on the left and on the right by $\hat{V}$ and taking account of (15) entail

$$\hat{\Sigma} = \hat{V} + (\hat{c}^t W\hat{c})\hat{V}\hat{b}\hat{b}^t\hat{V}. \quad (19)$$

As a consequence of (19), we have

$$\hat\Sigma\hat b = \Big(1 + (\hat c^t W \hat c)(\hat b^t \hat V \hat b)\Big)\, \hat V \hat b, \tag{20}$$

which means that $\hat\Sigma\hat b$ is proportional to $\hat V \hat b$. Consequently, one also has

$$\hat V \hat b = \theta(\hat b)\hat\Sigma\hat b, \tag{21}$$

where we have defined

$$\theta(\hat b) = \frac{\hat b^t \hat V \hat b}{\hat b^t \hat\Sigma \hat b}.$$

Substituting (21) in (18) yields

$$(\hat c^t W \hat c)(\hat b^t \hat V \hat b)\theta(\hat b)\hat\Sigma\hat b = M^t W^{-1} M \hat b$$

and thus b is an eigenvector of $\hat\Sigma^{-1} M^t W^{-1} M$. Let us denote by $\hat\lambda$ the associated eigenvalue

$$\hat\lambda = (\hat c^t W \hat c)(\hat b^t \hat V \hat b)\theta(\hat b). \tag{22}$$

Collecting (20) and (21) yields

$$\frac{1}{\theta(\hat b)} = 1 + (\hat c^t W \hat c)(\hat b^t \hat V \hat b),$$

and thus the eigenvalue can be rewritten as

$$\hat\lambda = 1 - \theta(\hat b).$$

Moreover, we have, from (16),

$$\begin{aligned}
\hat c^t M \hat b &= \hat\lambda/\theta(\hat b), && (23)\\
\operatorname{tr}(\hat\Sigma \hat V^{-1}) &= p + \hat\lambda/\theta(\hat b), && (24)\\
\log\det \hat V &= \log\det\hat\Sigma - \log\det\Big(I_p + (\hat c^t W \hat c)\hat V \hat b \hat b^t\Big) \\
&= \log\det\hat\Sigma - \log\Big(1 + \hat\lambda/\theta(\hat b)\Big), && (25)
\end{aligned}$$

entailing

$$\begin{aligned}
G(\hat\mu, \hat V, \hat b, \hat c) &= p + \log\det\hat\Sigma - \log\Big(1 + \hat\lambda/\theta(\hat b)\Big) \\
&= p + \log\det\hat\Sigma + \log(1 - \hat\lambda).
\end{aligned}$$

As a consequence, to minimize G, $\hat\lambda$ should be the largest eigenvalue. Finally, let us consider

$$V_0 = \hat\Sigma - \frac{\hat\lambda}{\hat b^t \hat\Sigma \hat b}\hat\Sigma \hat b \hat b^t \hat\Sigma,$$

leading to

$$\begin{aligned}
V_0 + (\hat c^t W \hat c) V_0 \hat b \hat b^t V_0 &= \hat\Sigma + \left((\hat c^t W \hat c)\theta^2(\hat b) - \frac{\hat\lambda}{\hat b^t \hat\Sigma \hat b}\right)\hat\Sigma \hat b \hat b^t \hat\Sigma \\
&= \hat\Sigma + \frac{\hat\lambda\theta(\hat b)}{\hat b^t V_0 \hat b}\Big((\hat c^t W \hat c)(\hat b^t V_0 \hat b)\theta(\hat b) - \hat\lambda\Big)\hat\Sigma \hat b \hat b^t \hat\Sigma \\
&= \hat\Sigma,
\end{aligned}$$

in view of (22) and thus V_0 verifies equation (19). ■

Proof of Corollary 1. Let us remark that, under (7), the coefficients of W_{ij} of W are given by $W_{ij} = f_i \mathbb{I}\{i = j\} - f_i f_j$ for all $(i,j) \in \{1,\dots,h\}^2$. The inverse matrix of W is

$$W^{-1} = \text{diag}\left(\frac{1}{f_1},\dots,\frac{1}{f_h}\right) + \frac{1}{f_{h+1}} U,$$

where U is the $h \times h$ matrix defined by $U_{ij} = 1$ for all $(i,j) \in \{1,\dots,h\} \times \{1,\dots,h\}$. Since the jth line of M is given by $f_j(\bar{X}_j - \bar{X})^t$ for all $j = 1,\dots,h$ and taking account of $U^2 = f_{h+1}U$, we have

$$\begin{aligned} M^t W^{-1} M &= \sum_{j=1}^{h} f_j (\bar{X}_j - \bar{X})(\bar{X}_j - \bar{X})^t + \frac{1}{f_{h+1}} M^t U M \\ &= \sum_{j=1}^{h} f_j (\bar{X}_j - \bar{X})(\bar{X}_j - \bar{X})^t + \frac{1}{h f_{h+1}} (M^t U)(M^t U)^t. \end{aligned} \tag{26}$$

Now, remarking that all the columns of $M^t U$ are equal to

$$\begin{aligned} \sum_{j=1}^{h} f_j (\bar{X}_j - \bar{X}) &= \sum_{j=1}^{h} f_j (\bar{X}_j - \bar{X}) - f_{h+1}(\bar{X}_{h+1} - \bar{X}) \\ &= -f_{h+1}(\bar{X}_{h+1} - \bar{X}), \end{aligned}$$

it follows that

$$(M^t U)(M^t U)^t = h f_{h+1}^2 (\bar{X}_{h+1} - \bar{X})(\bar{X}_{h+1} - \bar{X})^t$$

and thus replacing in (26) yields

$$\begin{aligned} M^t W^{-1} M &= \sum_{j=1}^{h+1} f_j (\bar{X}_j - \bar{X})(\bar{X}_j - \bar{X})^t \\ &= \hat{\Gamma}. \end{aligned}$$

The result is then a consequence of Proposition 1. ■

Proof of Lemma 2. The joint distribution of (X, b) given Y denoted by $p(X, b|Y)$ is calculated as the product $p(X|Y, b)p(b|Y)$ where $p(X|Y, b)$ is given by (4) and $p(b|Y)$ is given by (8). The estimators are obtained by minimizing

$$\begin{aligned} J_\Omega(\mu, V, b, c) &= -\frac{2}{n}\sum_{i=1}^{n} \log p(X_i, b|Y_i) = -\frac{2}{n}\sum_{i=1}^{n} \log p(X_i|Y_i, b) - \frac{2}{n}\sum_{i=1}^{n} \log p(b|Y_i) \\ &= G(\mu, V, b, c) + \frac{b^t \Omega^{-1} b}{1+\rho} \frac{1}{n} \sum_{i=1}^{n} ((s(Y_i) - \bar{s})^t c)^2 + C^{ste} \\ &= G(\mu, V, b, c) + \frac{b^t V b}{b^t \Sigma b} (b^t \Omega^{-1} b)(c^t W c) + C^{ste}, \end{aligned}$$

which is $G_\Omega(\mu, V, b, c)$ up to the constant C^{ste}. ■

Proof of Proposition 2. The proof is similar to the one of Proposition 1. First, remark that equation (11) still holds and thus $\hat{\mu} = \bar{X} - \bar{s}^t \hat{c} \hat{V} \hat{b}$. Let us recall the following definitions

$$\theta(\hat{b}) = \frac{\hat{b}^t \hat{V} \hat{b}}{\hat{b}^t \hat{\Sigma} \hat{b}} \text{ and } \eta(\hat{b}) = \frac{\hat{b}^t \Omega^{-1} \hat{b}}{\hat{b}^t \hat{\Sigma} \hat{b}}.$$

Annulling the gradients of $G_\Omega(\mu, V, b, c)$ yields the system of equations

$$\begin{aligned} (\hat{c}^t W \hat{c}) \left(\Omega \hat{V} \hat{b} + \theta(\hat{b}) \hat{b} + \eta(\hat{b}) \Omega (\hat{V} \hat{b} - \theta(\hat{b}) \hat{\Sigma} \hat{b}) \right) &= \Omega M^t \hat{c}, && (27) \\ (\hat{b}^t \hat{V} \hat{b})(1 + \eta(\hat{b})) W \hat{c} &= M \hat{b}, && (28) \\ \hat{V}^{-1} + \hat{b} \hat{b}^t (\hat{c}^t W \hat{c})(1 + \eta(\hat{b})) &= \hat{V}^{-1} \hat{\Sigma} \hat{V}^{-1}. && (29) \end{aligned}$$

Assuming W is regular, equation (28) entails

$$\hat{c} = W^{-1} M \hat{b} / ((1 + \eta(\hat{b}))(\hat{b}^t \hat{V} \hat{b}))$$

and replacing in (27) yields

$$(\hat{c}^t W \hat{c})(\hat{b}^t \hat{V} \hat{b})(1 + \eta(\hat{b})) \left(\Omega \hat{V} \hat{b} + \theta(\hat{b}) \hat{b} + \eta(\hat{b}) \Omega (\hat{V} \hat{b} - \theta(\hat{b}) \hat{\Sigma} \hat{b}) \right) = \Omega M^t W^{-1} M \hat{b}. \quad (30)$$

Now, multiplying (29) on the left and on the right by $\hat{V}$, it follows that

$$\hat{\Sigma} = \hat{V} + (\hat{c}^t W \hat{c})(1 + \eta(\hat{b})) \hat{V} \hat{b} \hat{b}^t \hat{V},$$

leading to

$$\hat{\Sigma} \hat{b} = \left(1 + (\hat{c}^t W \hat{c})(\hat{b}^t \hat{V} \hat{b})(1 + \eta(\hat{b})) \right) \hat{V} \hat{b}, \quad (31)$$

which means that $\hat{\Sigma} \hat{b}$ is proportional to $\hat{V} \hat{b}$. As a consequence, one also has

$$\hat{V} \hat{b} = \theta(\hat{b}) \hat{\Sigma} \hat{b}, \quad (32)$$

and replacing in (30) yields

$$(\hat{c}^t W \hat{c})(\hat{b}^t \hat{V} \hat{b})(1 + \eta(\hat{b})) \theta(\hat{b}) \left(\Omega \hat{\Sigma} + I_p \right) \hat{b} = \Omega M^t W^{-1} M \hat{b},$$

and thus b is an eigenvector of $(\Omega \hat{\Sigma} + I_p)^{-1} \Omega M^t W^{-1} M$. Let us denote by $\hat{\lambda}$ the associated eigenvalue

$$\hat{\lambda} = (\hat{c}^t W \hat{c})(\hat{b}^t \hat{V} \hat{b})(1 + \eta(\hat{b})) \theta(\hat{b}).$$

Collecting (31) and (32) yields

$$\frac{1}{\theta(\hat{b})} = 1 + (\hat{c}^t W \hat{c})(\hat{b}^t \hat{V} \hat{b})(1 + \eta(\hat{b}))$$

and consequently the eigenvalue can be rewritten as

$$\hat{\lambda} = 1 - \theta(\hat{b}).$$

Now, let us remark that (23), (24) and (25) still hold in this context entailing

$$G_\Omega(\hat{\mu}, \hat{V}, \hat{b}, \hat{c}) = p + \log \det \hat{\Sigma} + \log(1 - \hat{\lambda}).$$

As a consequence, to minimize G_Ω, $\hat{\lambda}$ should be the largest eigenvalue. The end of the proof follows the same lines as the one of Proposition 1. ■

Proof of Proposition 3. Let P be the projection matrix on S_d:

$$P = \sum_{j=1}^{d} q_j q_j^t,$$

and, for all $i = 1, \dots, n$ consider the projected predictor defined by $\tilde{X}_i = PX_i$. Introducing $\tilde{\Gamma} = P\hat{\Gamma}P$ the empirical "between slices" matrix associated to $\tilde{X}_1, \dots, \tilde{X}_n$ and $\tilde{\Sigma} = P\hat{\Sigma}P$ the corresponding covariance matrix, the PCA+SIR method finds $\tilde{b}$ such that

$$\tilde{\Gamma}\tilde{b} = \tilde{\lambda}\tilde{\Sigma}\tilde{b},$$

where $\tilde{\lambda} \in \mathbb{R}$, or equivalently,

$$P\hat{\Gamma}P\tilde{b} = \tilde{\lambda}P\hat{\Sigma}P\tilde{b}.$$

Remarking that $P\hat{\Sigma}P = \hat{\Sigma}P$, we have, for all $\tau > 0$,

$$P\hat{\Gamma}P\tilde{b} = \frac{\tilde{\lambda}}{1+\tau}(P\hat{\Sigma} + \tau\hat{\Sigma})P\tilde{b},$$

and defining $b = P\tilde{b}$ and $\lambda = \tilde{\lambda}/(1+\tau)$, it follows that

$$P\hat{\Gamma}b = \lambda(P\hat{\Sigma} + \tau\hat{\Sigma})b.$$

Since $P = \tau\hat{\Sigma}\Omega_2$, we thus have

$$\hat{\Sigma}\Omega_2\hat{\Gamma}b = \lambda(\hat{\Sigma}\Omega_2\hat{\Sigma} + \hat{\Sigma})b,$$

which means that b is an eigenvector of $(\Omega_2\hat{\Sigma} + I_p)^{-1}\Omega_2\hat{\Gamma}$. Corollary 2 concludes the proof, *i.e.* b is the GRSIR estimator with prior covariance matrix Ω_2. ■

Proof of Proposition 4. Here, we adopt the notations introduced in the proof of Proposition 3. The GRSIR estimator $\tilde{b}$ computed on the projected predictors $\tilde{X}_1, \dots, \tilde{X}_n$ verifies

$$\tilde{\Omega}(\varphi)\tilde{\Gamma}\tilde{b} = \tilde{\lambda}(\tilde{\Omega}(\varphi)\tilde{\Sigma} + I_p)\tilde{b},$$

or equivalently

$$\tilde{\Omega}(\varphi)P\hat{\Gamma}P\tilde{b} = \tilde{\lambda}(\tilde{\Omega}(\varphi)P\hat{\Sigma}P + I_p)\tilde{b}.$$

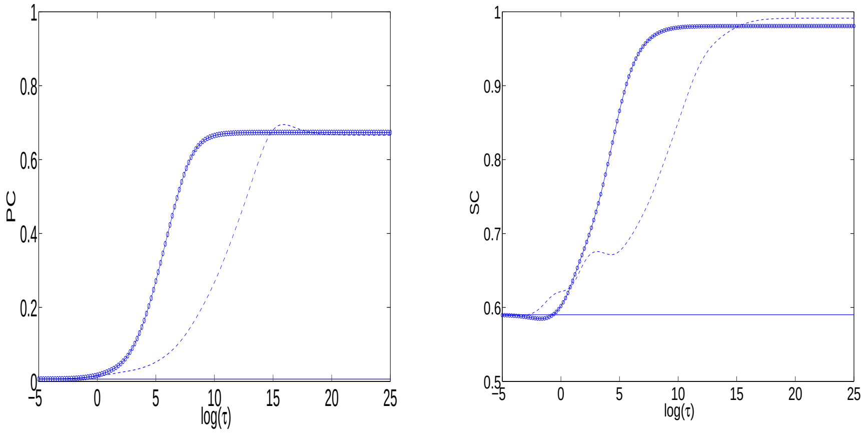
Multiplying this equation by P on the left, we obtain

$$P\tilde{\Omega}(\varphi)P\hat{\Gamma}P\tilde{b} = \tilde{\lambda}(P\tilde{\Omega}(\varphi)P\hat{\Sigma}P + P)\tilde{b}.$$

Since $P\tilde{\Omega}(\varphi)P = \Omega(\varphi)$, and introducing $b = P\tilde{b}$, it follows that

$$\Omega(\varphi)\hat{\Gamma}b = \tilde{\lambda}(\Omega(\varphi)\hat{\Sigma} + I_p)b,$$

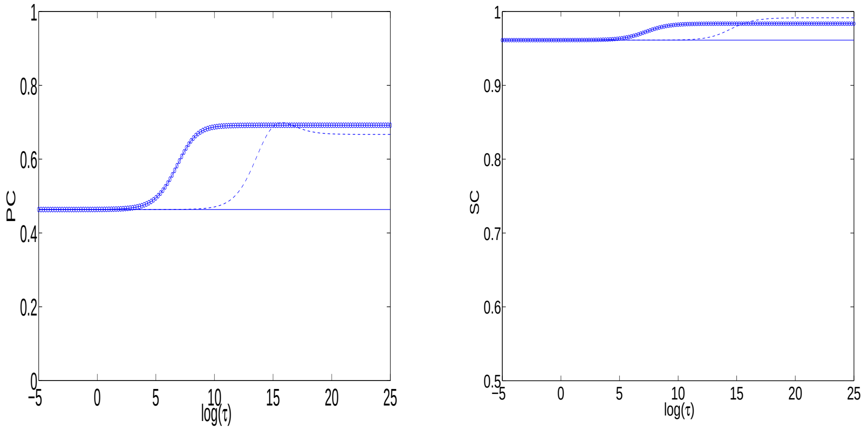
which means that b is an eigenvector of $(\Omega(\varphi)\hat{\Sigma} + I_p)^{-1}\Omega(\varphi)\hat{\Gamma}$. Corollary 2 concludes the proof, *i.e.* b is the GRSIR estimator with prior covariance matrix $\Omega(\varphi)$. ■



(a) Criterion MSC

(b) Criterion VSC

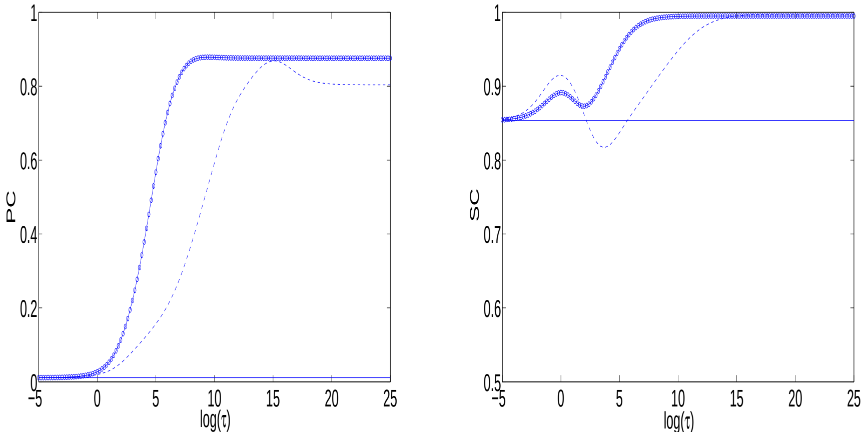
Figure 1: Values of MSC and SSC with respect to the regularization parameter for Model 1. The condition number is fixed to $\theta = 2$. Horizontally: $\log(\tau)$, vertically: (a) MSC and (b) VSC. Continuous line: Ω_0 (SIR), "-o-" line: Ω_1 (ridge), dashed line: Ω_3 (Tikhonov).



(a) Criterion MSC

(b) Criterion VSC

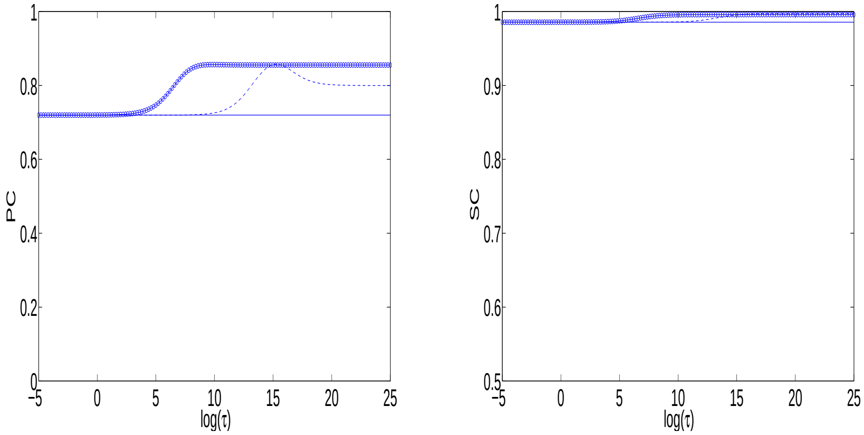
Figure 2: Values of MSC and VSC with respect to the regularization parameter for Model 1. The cut-off dimension is chosen to $d = 20$ and the condition number is fixed to $\theta = 2$. Horizontally: $\log(\tau)$, vertically: (a) MSC and (b) VSC. Continuous line: Ω_2 (PCA+SIR), "-o-" line: Ω_4 (PCA+ridge), dashed line: Ω_5 (PCA+Tikhonov).



(a) Criterion MSC

(b) Criterion VSC

Figure 3: Values of MSC and VSC with respect to the regularization parameter for Model 2. The condition number is fixed to $\theta = 2$. Horizontally: $\log(\tau)$, vertically: (a) MSC and (b) VSC. Continuous line: Ω_0 (SIR), ”-o-” line: Ω_1 (ridge), dashed line: Ω_3 (Tikhonov).



(a) Criterion MSC

(b) Criterion VSC

Figure 4: Values of MSC and VSC with respect to the regularization parameter for Model 2. The cut-off dimension is chosen to $d = 20$ and the condition number is fixed to $\theta = 2$. Horizontally: $\log(\tau)$, vertically: (a) MSC and (b) VSC. Continuous line: Ω_2 (PCA+SIR), ”-o-” line: Ω_4 (PCA+ridge), dashed line: Ω_5 (PCA+Tikhonov).

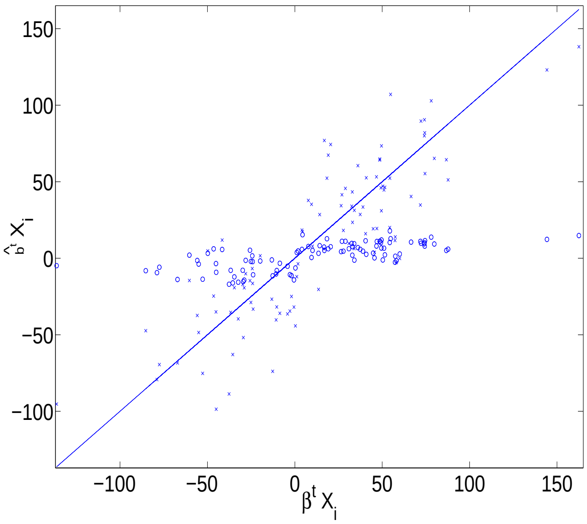


Figure 5: An example of pairs $(\beta^t X_i, \hat{b}^t X_i)$, $i = 1, \ldots, n$ for Model 1 where the axis $\hat{b}$ is computed with the classical SIR method (points "o") and the PCA+Tikhonov method (points "x"). Here, $\theta = 2$ and $d = 20$.

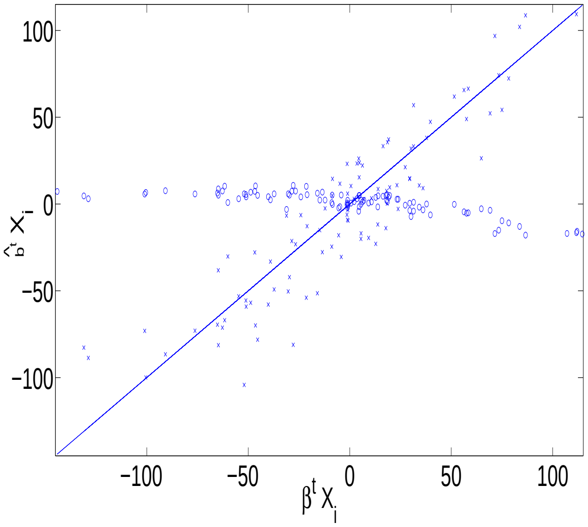


Figure 6: An example of pairs $(\beta^t X_i, \hat{b}^t X_i)$, $i = 1, \ldots, n$ for Model 2 where the axis $\hat{b}$ is computed with the classical SIR method (points "o") and the PCA+Tikhonov method (points "x"). Here, $\theta = 2$ and $d = 20$.

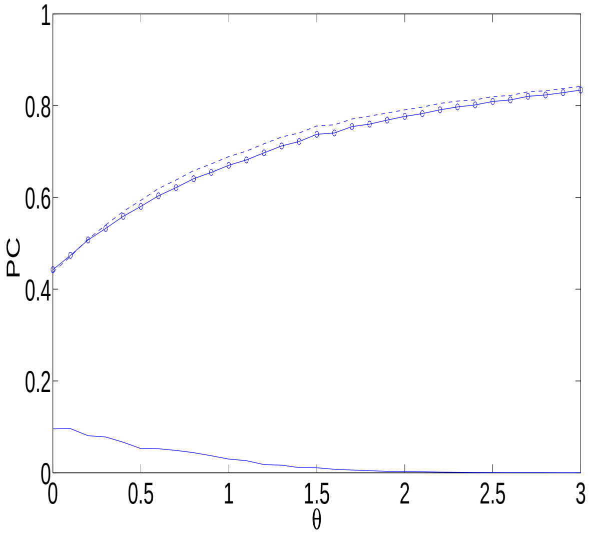


Figure 7: Sensibility of GRSIR with respect to the condition number of the covariance matrix for Model 1. Here, we use the optimal τ. Horizontally: θ, vertically: the criterion MSC. Continuous line: Ω_0 (SIR), "-o-" line: Ω_1 (ridge), dashed line: Ω_3 (Tikhonov).

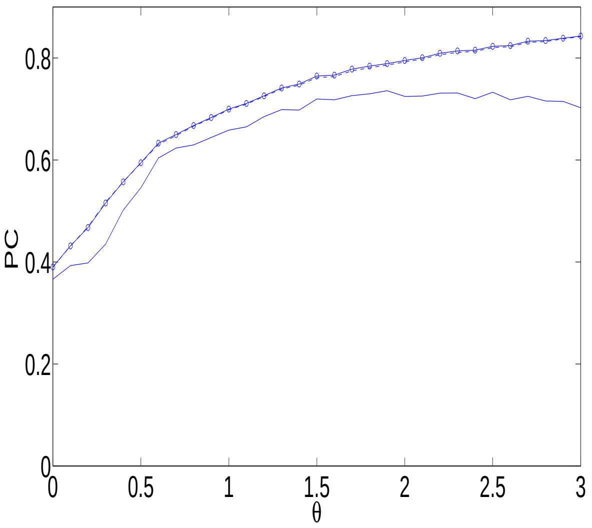


Figure 8: Sensibility of GRSIR with respect to the condition number of the covariance matrix for Model 1. The cut-off dimension is chosen to $d = 20$ and we use the optimal τ.. Horizontally: θ, vertically: the criterion MSC. Continuous line: Ω_2 (PCA+SIR), "-o-" line: Ω_4 (PCA+ridge), dashed line: Ω_5 (PCA+Tikhonov).

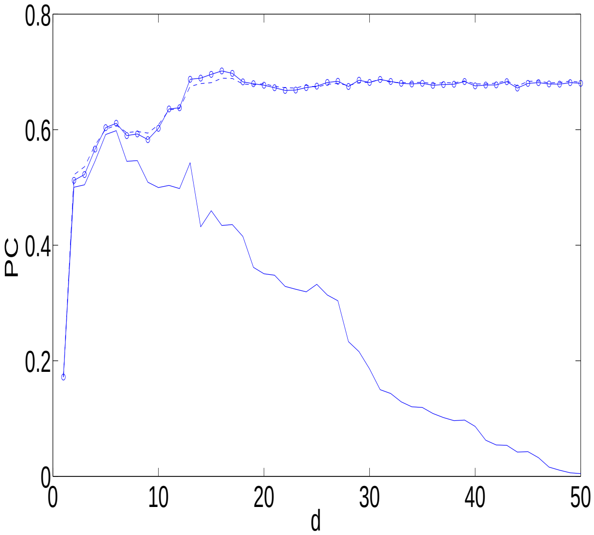


Figure 9: Sensibility of GRSIR with respect to the cut-off dimension for Model 1. Horizontally: d, vertically: the MSC criterion. Continuous line: Ω_2 (PCA+SIR), "-o-" line: Ω_4 (PCA+ridge), dashed line: Ω_5 (PCA+Tikhonov). Here, $\theta = 2$ and the optimal τ is used.

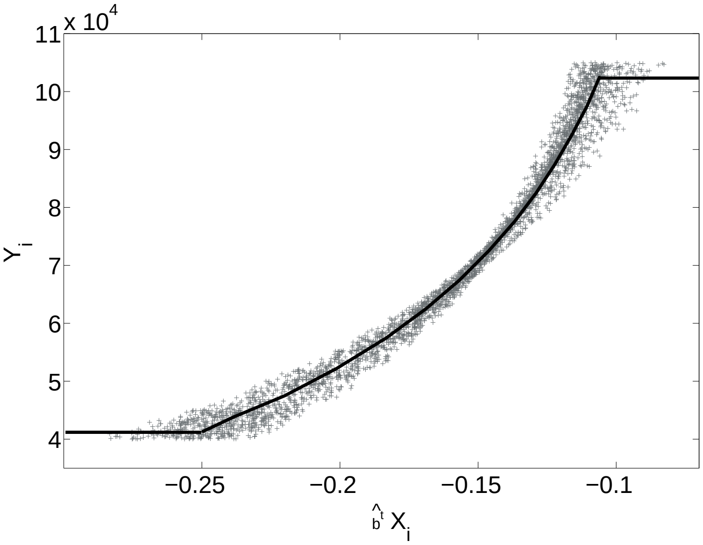


Figure 10: Functional relationship between reduced spectra $\hat{\beta}^t X$ on the first GRSIR direction and Y, the grain size of CO_2. Horizontally: reduced spectra from the learning database on the first GRSIR direction. Vertically: Grain size of CO_2.

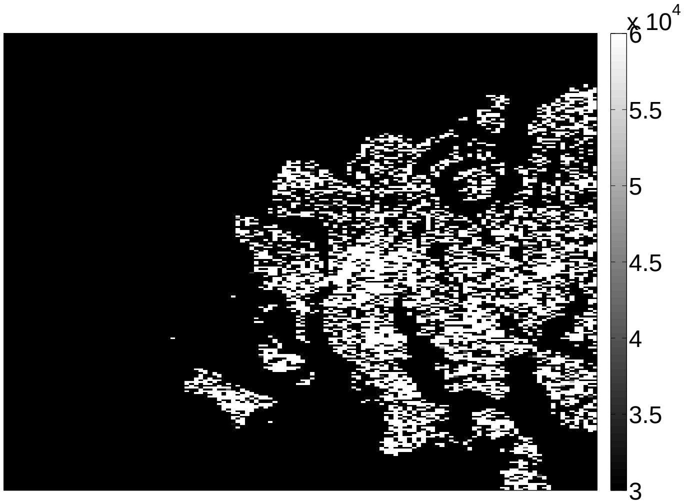


Figure 11: Grain size of CO_2 ice estimated by SIR on an hyperspectral image observed on Mars during orbit 61.

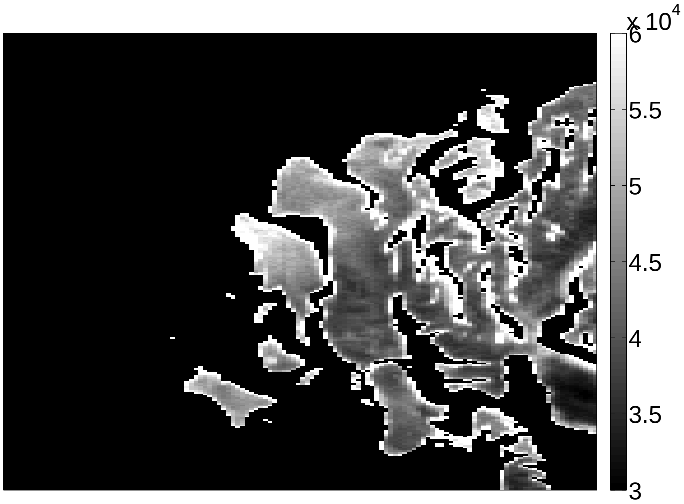


Figure 12: Grain size of CO_2 ice estimated by GRSIR (PCA+ridge) on an hyperspectral image observed on Mars during orbit 61.